WISERD
Sefydliad Ymchwil Cymdeithasol ac Economaidd a Data Cymru
Wales Institute of Social and Economic Research and Data

BRITISH ELECTION STUDY

Does ChatGPT Know What the Most Important Issue is?

Using Large Language Models to Code Open-Text Social Survey Responses at Scale

Dr Ralph Scott

Co-authored with Dr Jonathan Mellon, Dr Jack Bailey, James Breckwoldt and Marta Miori

Research funded by

UKRI Economic and Social Research Council

Today's structure

- Setting out the research problem
- Introducing ChatGPT (and LLMs generally)
- How we tested its potential and what we found
- Potential and issues for future research
- Q&A

- Caveat: we approach this as social scientists, not computer scientists!

TL; DR – summary

- State of the art LLMs appear capable of performing open-text survey labelling tasks at close to human-level performance.
  - GPT-3 is able to match the original human coder's collapsed category 95% of the time, similar to human performance and better than the SVM.
- This is particularly impressive given the lack of training data provided.
- Where it was wrong, this was often relating to edge cases that human coders might also disagree on – we can also learn from LLMs in coding!
- Yet performance was weaker compared with the human coder on more complex tasks – this may improve as LLMs develop further.

The research problem

- Open-text survey responses offer certain advantages.
  - For instance, they avoid priming respondents to give particular answers (Ferrario and Stantcheva 2022; Esses and Maio 2002).
  - They also do not prejudge the answers that survey respondents might give (Geer 1991; Schuldt and Roh 2014).

- But they also have one significant drawback: they are costly and time-consuming to code for further analysis.
  - It would be better to automate this process, but historically this has not reliably achieved human coder standards.
  - However, can new tools like ChatGPT code open text well enough to use in public opinion research?

The British Election Study

- The British Election Study is one of the longest running election studies world-wide and the longest running social science survey in the UK.
- Surveys have taken place immediately after every general election since 1964, and the current study has been running since 2014.
- So far there have been 25 waves of the internet panel, each with around 30,000 respondents.
- In each wave we ask the following question:

"As far as you're concerned, what is the SINGLE MOST important issue facing the country at the present time?"

BRITISH ELECTION STUDY

WISERD

The approach so far

- This most important issue (MII) question is a key indicator of issue salience (Dennison 2019; Bevan, Jennings, and Wlezien 2016).
- So far, MII responses have been labelled by human coders who assign responses to 50 granular and 13 collapsed categories.
- We have manually labelled over 657,000 open-text responses to the MII question in this way since 2014.
- If this can be automated reliably, it would free up time for more creative research activity, and make more open response questions a possibility.

The categories

	Collapsed		Full
1	Europe	15	Europe
2	Immigration	12	Immigration
3	Economy	26	Economy-general
		27	Economy-personal
		28	Unemployment
		29	Taxation
		30	Debt/deficit
		31	Inflation
		32	Living costs
4	Health	1	Health
		48	Coronavirus
		49	COVID-economy
5	Terrorism	11	Terrorism
7	Inequality	33	Poverty
		35	Inequality
		36	Housing
8	Environment	40	Environment
9	Austerity/spending	2	Education
		10	Welfare
		34	Austerity
		37	Social care
		38	Pensions/ageing
10	Negativity	4	Pol-neg
		5	Partisan-neg
		6	Societal divides
11	Other lib-auth	7	Morals
		8	Nat ident, goals loss
		9	Racism/discrimination
		14	Crime
		21	Foreign affairs
		22	War
		23	Defence
		41	Pol values-auth
		42	Pol values-liberal
		50	BLM and responses
12	Other left-right	43	Pol values-right
		44	Pol values-left
13	Other	16	Consitutional
		17	International trade
		18	Devolution
		19	Scot-ind
		24	Foreign emergency
		25	Domestic emergency
		45	Other
		47	Referendum unspecified
		46	Uncoded

The raw data

	A	B	C	D	E	F
37	12387	24	Alarmism	46	uncoded	
38	51518	24	All of it has gone to crap	46	uncoded	
39	83361	24	All of the strikes happening	26	economy-general	
40	17297	24	All the strikes	26	economy-general	
41	41434	24	all the strikes	26	economy-general	
42	47445	24	All the strikes	26	economy-general	
43	89465	24	all the strikes	26	economy-general	
44	106765	24	All the strikes, namely the NHS being underfunded	26	economy-general	
45	91437	24	Ambulance strikes	1	health	
46	5406	24	An Ageing population	38	pensions/ageing	
47	973	24	An incompetant government	4	pol-neg	
48	100821	24	An incompetent Government	4	pol-neg	
49	15553	24	An inept and failing government out of ideas	4	pol-neg	
50	57584	24	An inept Tory Government	5	partisan-neg	
51	107887	24	An out of control cost of living crisis.	32	living costs	
52	12250	24	Anarchy	46	uncoded	
53	57486	24	Animal Experiments In Laboratories	42	pol values-liberal	
54	15243	24	Animal welfare and animal abuse	42	pol values-liberal	
55	59407	24	Anthropogenic climate change	40	environment	
56	83774	24	Anti social behaviour	14	crime	
57	90419	24	Anti social behaviour	14	crime	

Introducing ChatGPT

- Released by OpenAI on 30 November 2022.
- Fastest-growing user base in internet history.
- Underlying model is OpenAI's GPT-3.5, fine-tuned to respond to human prompts (Ouyang et al 2022).
- A large language model (LLM) trained to predict the next most likely word that will occur after a user input (LeCun, Bengio, and Hinton 2015).
- Impressively, requires few or even zero in-context examples to achieve a good degree of accuracy (Brown et al 2020).
- In a state of constant development, GPT-4 has been recently released (and seems to perform even better!).

Example: original poetry

For instance, GPT-3 (accessed via chatGPT) can complete the prompt:

```
Write a 3 line poem about Margaret Atwood eating a canary
```

With a novel response that never appeared in its training data:

```
Margaret Atwood dines,
With a canary on her plate,
A fleeting, feathered feast.
```

Misusing ChatGPT

ChatGPT: US lawyer admits using AI for case research

27 May

REUTERS

ChatGPT can answer questions using natural, human-like language and mimic other writing styles

By Kathryn Armstrong

BBC News

A New York lawyer is facing a court hearing of his own after his firm used AI tool ChatGPT for legal research.

A judge said the court was faced with an "unprecedented circumstance" after a filing was found to reference example legal cases that did not exist.

The lawyer who used the tool told the court he was "unaware that its content could be false".

ChatGPT creates original text on request, but comes with warnings it can "produce inaccurate information".

BRITISH ELECTION STUDY

WISERD

Our approach: data

- Our test case is the 81,266 open text responses given to the MII question in waves 21-23 of the BES panel (May 2021-May 2022).
- These have already been hand-coded by a research assistant, providing a "correct" answer to compare predicted labels against.
- We assess accuracy across two samples (always the same across coders):
  - 1000 randomly sampled open responses
  - 1000 randomly sampled unique open responses
- This latter should be a harder test, because it will overrepresent rare and idiosyncratic responses.

Our approach: methods

- We compare how well GPT-3 codes open text responses to two relevant alternatives: human coding and supervised learning (using SVMs).
- Human:
  - A public opinion specialist who received an hour of training with examples drawn from a separate sample to the test data, and feedback from an experienced coder.
- SVM:
  - Provides a baseline for the performance of other machine learning methods.
  - We fit one SVM to 1,000 randomly sampled responses from waves 21-23 of BESIP.
    - Use case where a researcher labels a moderate sample of training data, then uses a supervised learning algorithm to label a much larger dataset.
  - We also fit a second SVM to ~576,000 responses from waves 1-20 of BESIP.
    - Use case of fitting a model to a large collection of existing coding and applying it forwards, an approach that will often not be feasible, but which therefore provides a strong test.

The GPT prompt

- To construct an appropriate prompt for GPT-3, we first experimented with ChatGPT to understand the capabilities and limits of the engine.
- Our prompt starts by saying:

```
Here are some open-ended responses from the British Election
Study to the question "what is the most important issue facing
the country?". Please assign one of the following categories to
each open ended text response, returning the original response
and the most relevant numeric code.
```

- We then listed the possible categories.

The GPT prompt

- We found that conversationally correcting the AI on its approach worked well, but that we could also pre-empt errors by adding some detail.

```
pol-neg: complaints about politics, the system, the media, corruption
where no politician or party is mentioned

europe: including Brexit
```

- We additionally informed the AI about Russia's invasion of Ukraine which took place after GPT-3's training data:

```
For context, Russia invaded Ukraine prior to the fieldwork for
this survey, so references to Ukraine or Russia are about war.
```

The GPT prompt

- We then provided two examples of the response format:

```
Code these cases:
a bad economy
Immergration

a bad economy|economy-general
immergration|immigration
```

The GPT prompt

- And then one more example correcting an error that was made in testing, where GPT returned two codes for a given response:

```
Code these cases:
climate change and unemployment

climate change and unemployment|environment,unemployment

Please only return one code per response (if multiple match use
the first issue listed). The correct response should be:

climate change and unemployment|environment
```

The GPT prompt

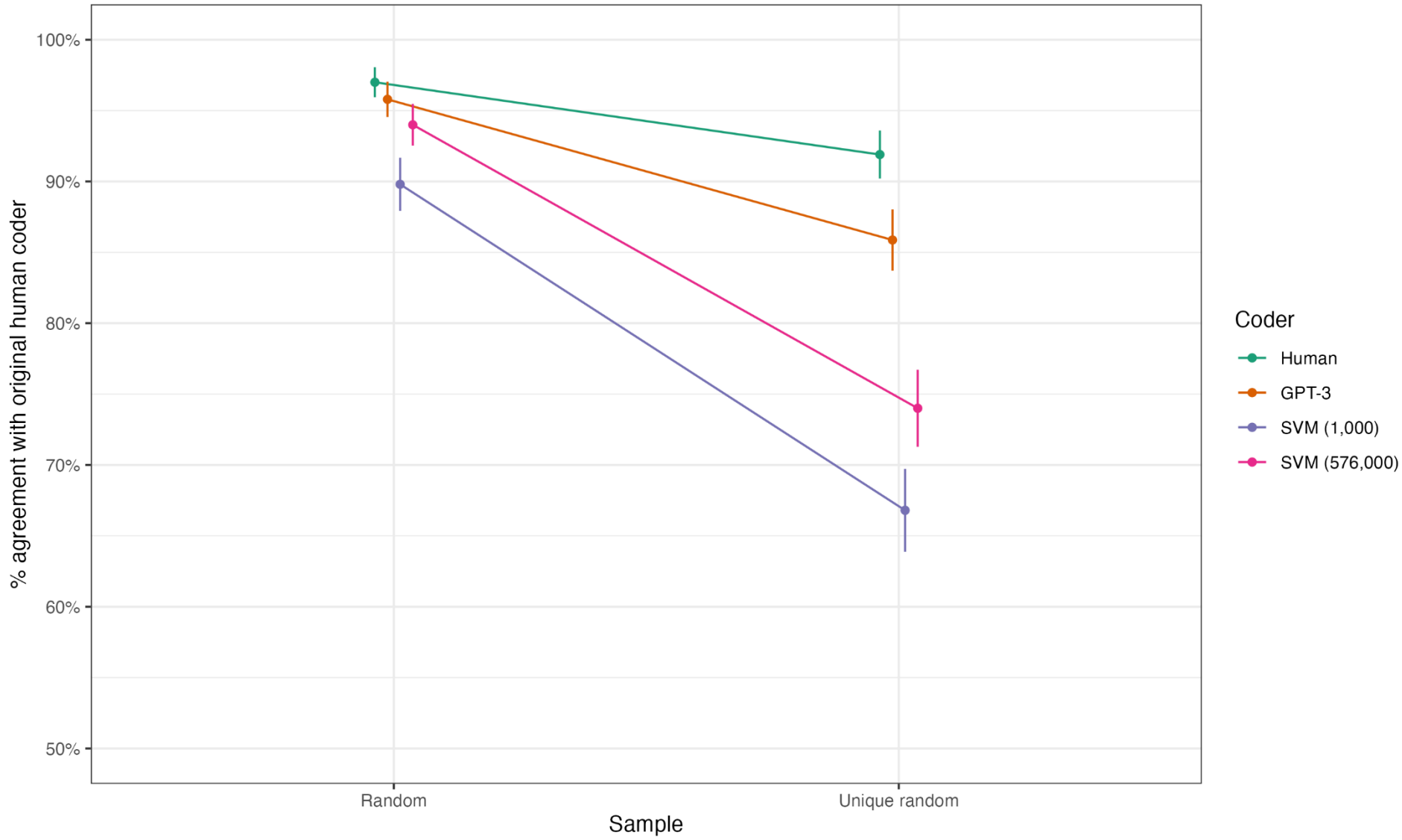
- We then finished by saying:

```
Code these cases:
```

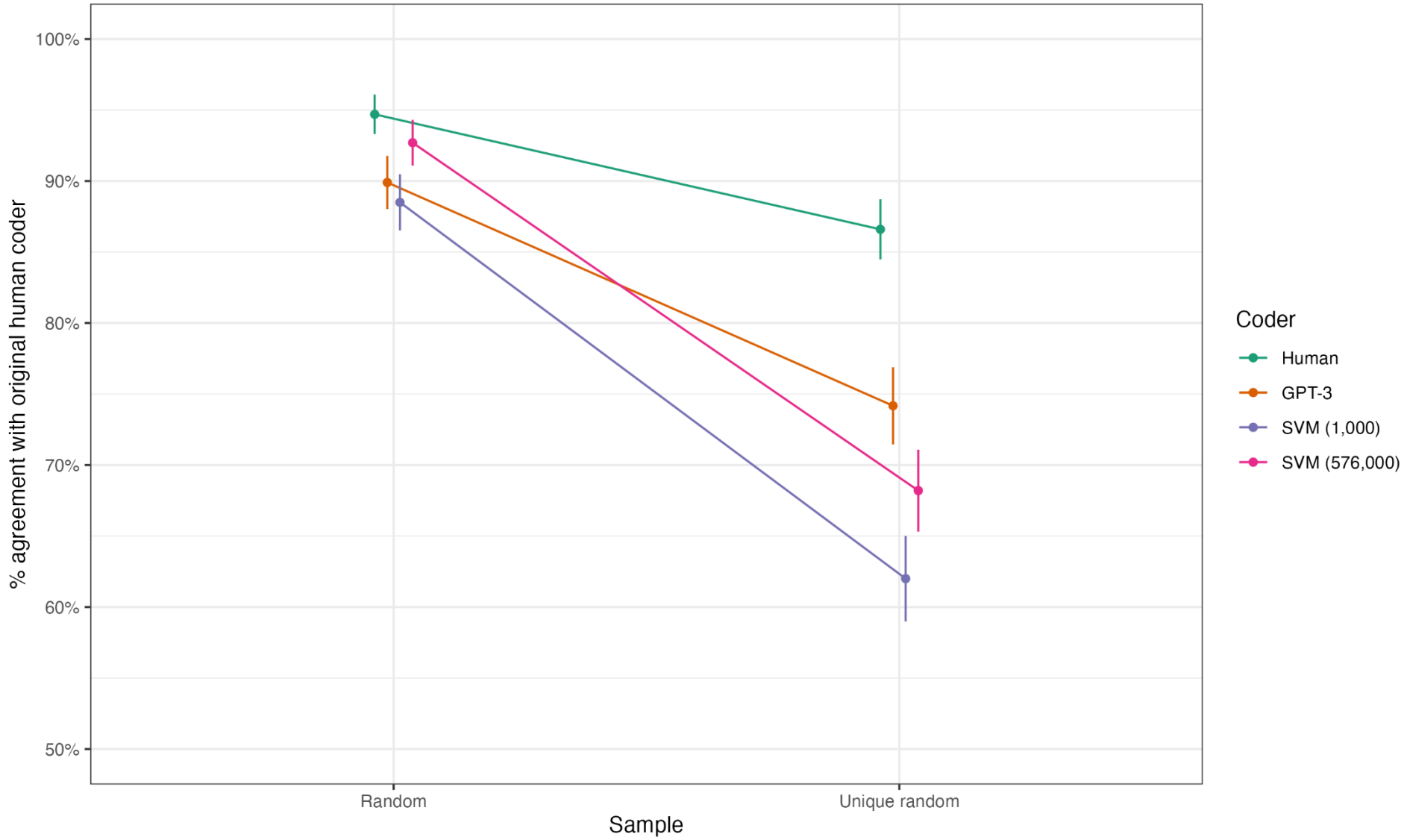
- And listed 50 open-ended responses at a time.

- All of this was input via R using Open AI's API.
- We experimented with different batch sizes and response formats, but found that this approach balanced efficiency and performance best.

Results: 13 categories



Results: 50 categories



What did GPT get wrong?

- The most common error was for GPT-3 to code a response as “covid-economy” when the original human coder coded it as “coronavirus”.

```
The after effects of the pandemic
The relaxing of the covid restrictions
Opening up the services sector
End lockdown
Post covid recovery
Reopening after Covid
```

- In most cases, either interpretation can be reasonably justified depending on how much economic subtext a coder is willing to infer from the text.

What did GPT get wrong?

- Another common category of coding differences was where GPT-3 gave the "uncoded" designation and the human coded the text as "coronavirus".

```
dumb people who think covid is a pandemic and do no research
Getting back to a form of normality
Pandamia
Covis
Lifting all restrictions
Omicron restrictions
```

- In this case, the human is more willing to make a reasonable inference that COVID is being referred to, while GPT-3 conservatively marks it uncoded.

What did GPT get wrong?

- A final common category of error was where GPT-3 labelled a text response as “war” while the human coder labelled it as “foreign affairs”.

```
Ensuryng that ?Ukraine does not lose to Adolf Putin
The crisis in ukraine
Putin’s madness
Russian hostility
Ukrainian crisis and threat of escalation from Russia
Threat to world peace by Russia
```

- Here the human coder has interpreted the war category relatively narrowly, and we have since actually adapted our coding schema based on this!

Findings

- State of the art LLMs such as GPT-3 appear capable of performing open-text survey labelling tasks at close to human-level performance.
  - GPT-3 is able to match the original human coder's collapsed category 95% of the time, similar to human performance and better than the SVM.
- This is particularly impressive given the lack of training provided: this performance is after only providing three examples in the prompt.
- Where it was wrong, this was often relating to edge cases that human coders might also disagree on – we can also learn from LLMs in coding!
- This task was likely a best-case scenario: text coding tasks that require more specialised knowledge will probably show lower performance.
- In addition, performance was weaker compared with the human coder on more complex tasks – this may improve as LLMs develop further.

Extensions

- More models:
  - GPT-4, GPT-3.5-turbo, Google's Bard and PaLM 2 and Anthropic's Claude all appear to be able to complete the task in some form.
  - We also tried current open source models (Vicuna, Meta's LLaMa and Alpaca, Google's FLAN) and none were able to complete the task.
- More performance metrics:
  - Beyond accuracy, including F1, Cohen's Kappa and ROC AUC.
- More use cases:
  - Eg coding of occupational classifications, NS-SEC.

Future potential and issues

- Potential:
  - Models are only going to get bigger and better.
  - This opens up new research possibilities: eg researchers who would prefer to use a different schema can now cheaply run their own classification on an LLM.
  - Cheap text coding makes the wider use of open-ended survey questions more viable, which could in turn provide answers we would never have expected.
- Issues:
  - Accuracy: not yet quite as good as human coding.
  - Accessibility: requires either an API or a powerful computer to be viable.
  - Interpretability: black box nature of the model means we don't know what exactly is driving performance, or what biases might be hidden.
  - Replicability and transparency: GPT-3.5 is privately owned and costs money to use. It could also change at any time. Are open-source models the solution?

WISERD

Sefydliad Ymchwil Cymdeithasol ac Economaidd a Data Cymru

Wales Institute of Social and Economic Research and Data

Thank you

Email: ScottR7@cardiff.ac.uk

Twitter: @ralphascott

Working paper: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4310154

Research funded by

UKRI Economic and Social Research Council

Constantly improving

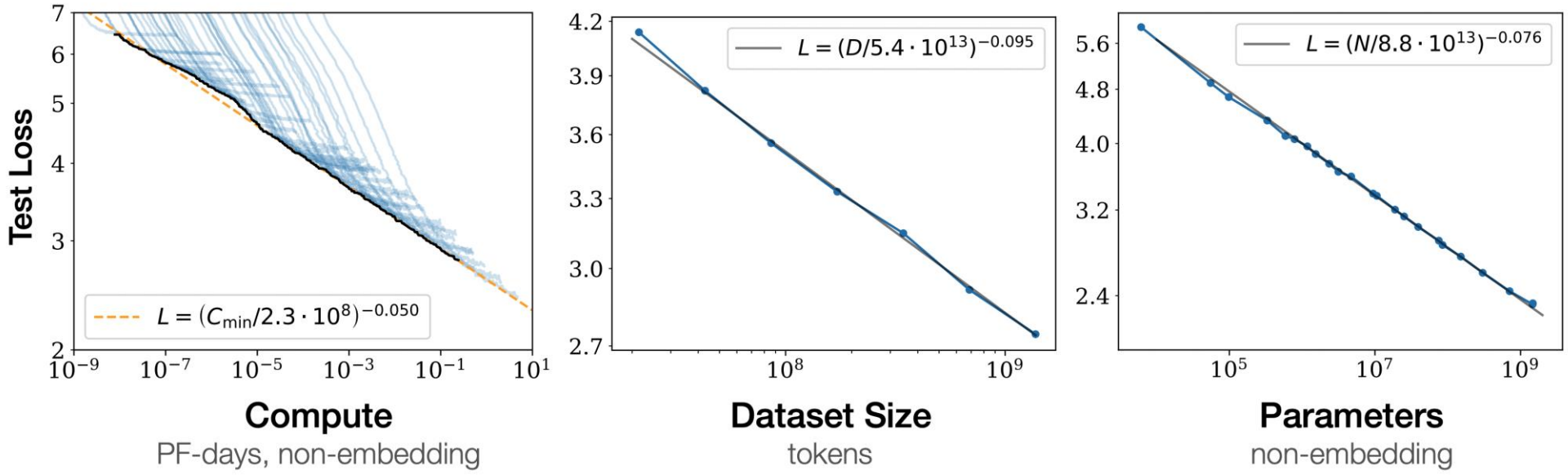


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

From Kaplan et al (2020)

Training ChatGPT

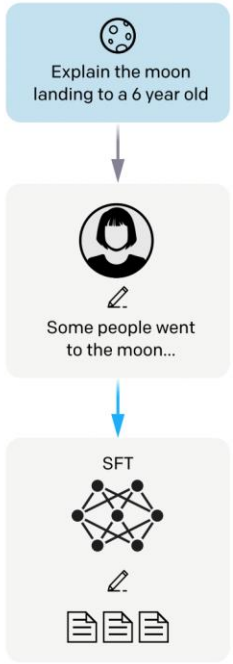
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



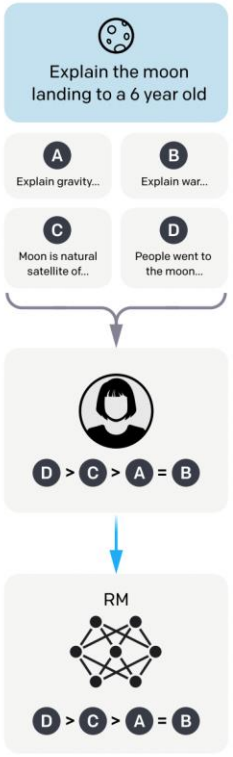
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

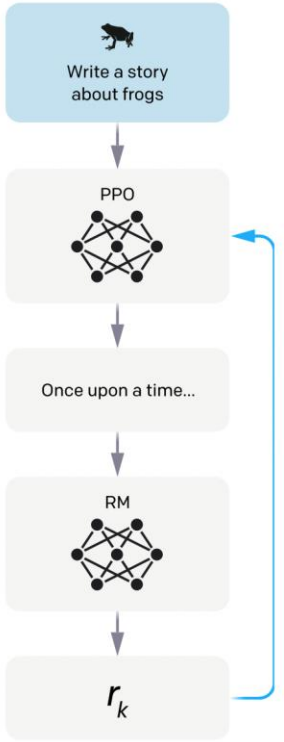
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



From Ouyang et al (2022)

BRITISH ELECTION STUDY

WISERD

Output issues

- GPT-3 was generally good about using only the categories given, but we had to make a few manual edits to its labels.
  - For instance, it used “covid” rather than “coronavirus” as a label in some cases and corrected the capitalization of the “europe” category to “Europe”.

- In addition, despite repeated instructions to return only a single code, GPT-3 still sometimes returned multiple codes.
  - We used only the first code returned in these cases.

Other models tried

Model	Creator	Version	Accessed through	Open-Source	Completes task	Follows format
GPT-4	Open AI	5/1/2023	nat.dev	FALSE	TRUE	TRUE
claude-v1	Anthropic	5/1/2023	nat.dev	FALSE	TRUE	TRUE
text-davinci-003	Open AI	5/1/2023	nat.dev	FALSE	TRUE	TRUE
text-davinci-002	Open AI	5/1/2023	nat.dev	FALSE	TRUE	TRUE
GPT-3.5-turbo	Open AI	5/1/2023	nat.dev	FALSE	TRUE	FALSE
Bard	Google	11:59 ET 5/1/2023	bard.google.com	FALSE	TRUE	FALSE
vicuna-13b	Meta/Stanford/Vicuna Team	a68b8408	replicate.com	TRUE	FALSE	FALSE
text-ada-001	Open AI	5/1/2023	nat.dev	FALSE	FALSE	FALSE
text-babbage-001	Open AI	5/1/2023	nat.dev	FALSE	FALSE	FALSE
text-curie-001	Open AI	5/1/2023	nat.dev	FALSE	FALSE	FALSE
xlarge	co:here	5/1/2023	nat.dev	FALSE	FALSE	FALSE
luminous-supreme-control	Aleph	5/1/2023	nat.dev	FALSE	FALSE	FALSE
llama13b	Meta	5/1/2023	nat.dev	TRUE	FALSE	FALSE
alpaca-7b	meta/Stanford	5/1/2023	nat.dev	TRUE	FALSE	FALSE
pythia-20b	Forefront	5/1/2023	nat.dev	TRUE	FALSE	FALSE
bloomz	bigscience	5/1/2023	nat.dev	TRUE	FALSE	FALSE
GPT-NeoX	Eleuther	5/1/2023	nat.dev	TRUE	FALSE	FALSE
flan-t5-xxl	Google	5/1/2023	nat.dev	TRUE	FALSE	FALSE
flan-ul2	Google	5/1/2023	nat.dev	TRUE	FALSE	FALSE
dolly-v2-12b	Databricks	ef0e1aef	replicate.com	TRUE	FALSE	FALSE
stalelm-tuned-alpha-7b	Stability AI	c49dae36	replicate.com	TRUE	FALSE	FALSE
llama-7b	meta	2014ee12	replicate.com	TRUE	FALSE	FALSE